

# NLP, Data Visualization, and the Mueller Report

---

Mike Mitri  
CIS & BSAN  
College of Business



# Overview

---

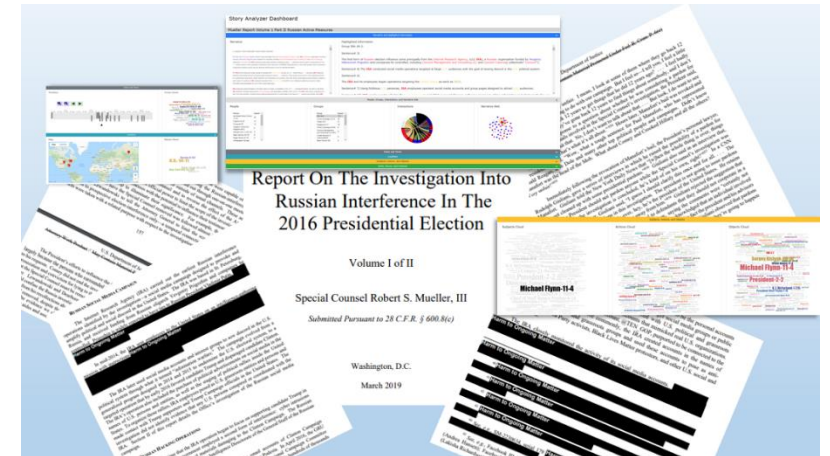
Story Analyzer <http://storyanalyzer.org/>

Natural Language Processing – Stanford’s CoreNLP API

Information Extraction

Data Visualization – Data Driven Documents (d3) and Google

“A picture tells a thousand words”



# Story Analyzer <http://storyanalyzer.org/>

---

An app that helps users visualize and understand a **story**

Story = a narrative

- ❑ People and groups are actors in the story. Actors can be subjects or objects.
- ❑ Subjects perform actions, objects receive the actions
- ❑ Times, places, and other contexts of actions
- ❑ Who did what to whom, where and when did it happen, and what else was going on at the time?

Story Analyzer uses NLP and data visualization APIs

Information extraction – take NLP results and capture narrative elements and relationships

Visualization – produce dashboards with visualizations depicting people, groups, interactions, times, and places

# LinkedIn Articles

---

[https://www.linkedin.com/posts/mike-mitri-a8912a1\\_activity-6580943797704810496-SAGg](https://www.linkedin.com/posts/mike-mitri-a8912a1_activity-6580943797704810496-SAGg)

[https://www.linkedin.com/posts/mike-mitri-a8912a1\\_datavizualization-datastorytelling-activity-6586329975928999936-es-](https://www.linkedin.com/posts/mike-mitri-a8912a1_datavizualization-datastorytelling-activity-6586329975928999936-es-)

[https://www.linkedin.com/posts/mike-mitri-a8912a1\\_nlp-congress-activity-6588102587873050624-hRV4](https://www.linkedin.com/posts/mike-mitri-a8912a1_nlp-congress-activity-6588102587873050624-hRV4)

[https://www.linkedin.com/posts/mike-mitri-a8912a1\\_nlp-datavizualization-datastorytelling-activity-6591323886409834496-W-nc](https://www.linkedin.com/posts/mike-mitri-a8912a1_nlp-datavizualization-datastorytelling-activity-6591323886409834496-W-nc)

[https://www.linkedin.com/posts/mike-mitri-a8912a1\\_nlp-storyanalyzer-informationextraction-activity-6595433941023420416-h953](https://www.linkedin.com/posts/mike-mitri-a8912a1_nlp-storyanalyzer-informationextraction-activity-6595433941023420416-h953)

# Natural Language Processing (NLP)

---

A mix of artificial intelligence and computational linguistics

The study of "understanding" the natural human language

NLP involves grammar (syntax) and semantics (meaning)

What is "Understanding"?

- ❑ Human understands, what about computers?
- ❑ Natural language is vague, context driven
- ❑ True understanding requires extensive knowledge of a topic. Beyond the scope of NLP itself...requires ontologies.

[http://en.wikipedia.org/wiki/Natural\\_language\\_processing](http://en.wikipedia.org/wiki/Natural_language_processing)

# Stanford's CoreNLP

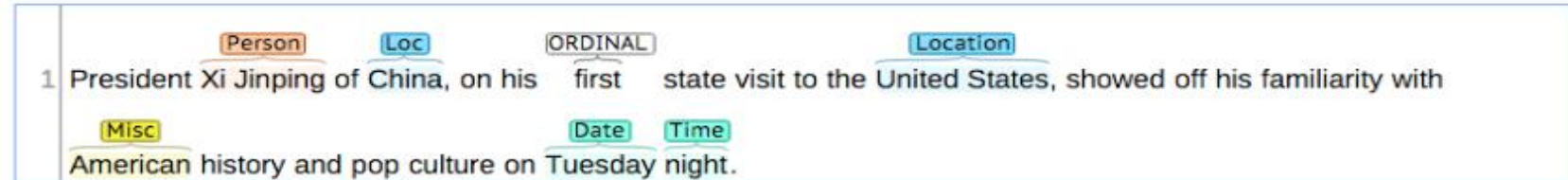
---

An open source Java-based API of classes and functions that can do several things:

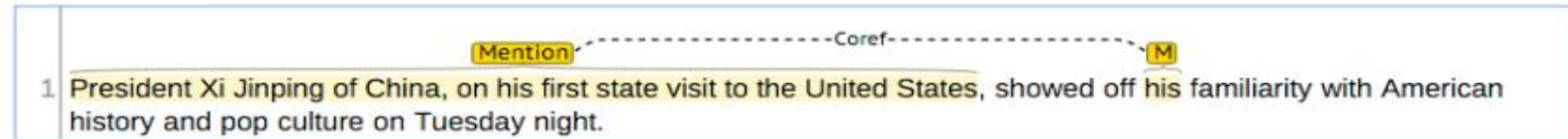
- Sentence splitting – breaking a text document into individual sentences
- Tokenizing a sentence (breaking it into individual “words”)
- Identifying parts of speech (POS) within a sentence (nouns, verbs, adjectives, adverbs, etc.)
- Named entity recognition: Recognizing names of people, places, organizations
- Constituency parsing
- Dependency parsing
- Co-reference resolution – finding all expressions that refer to the same entity in a text. (e.g. finding connections between nouns and their associated pronouns)
- Temporal tagging – recognizing and normalizing temporal expressions

# Stanford CoreNLP Features

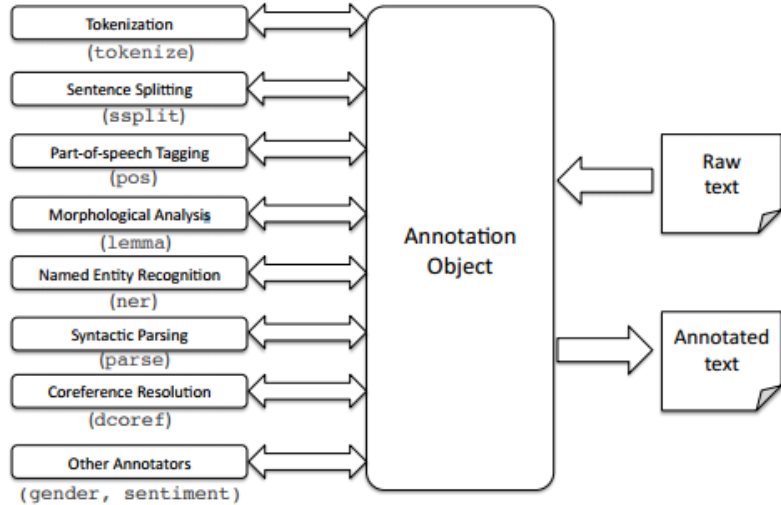
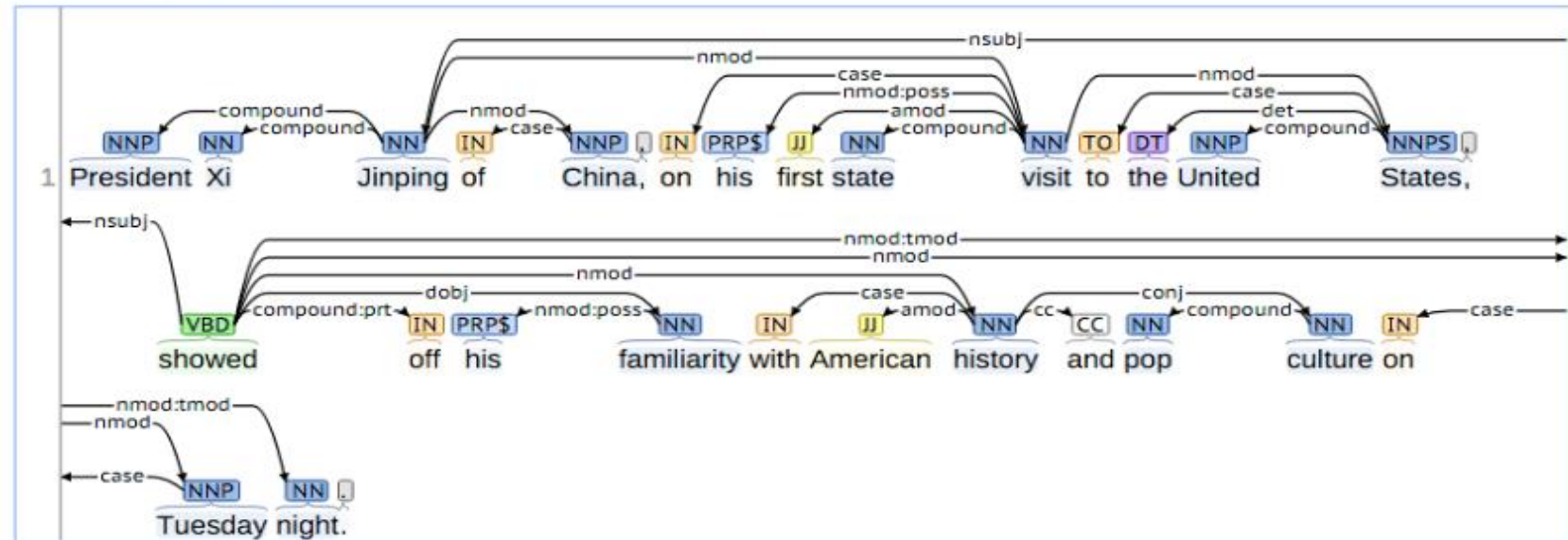
## Named Entity Recognition:



## Coreference:



## Basic Dependencies:



Some annotators based on machine learning, others rule-based.

# Sentence Splitting

---

Sentences end in periods, question marks, or exclamation points.

But just because you have a period doesn't mean you are at the end of a sentence:

- ❑ Mr. Jones
- ❑ Samantha G. Jones
- ❑ Here is some text (i.e. something written).
- ❑ A, B, C, etc., etc., etc.
- ❑ Go to website [cob.jmu.edu](http://cob.jmu.edu).
- ❑ What if the period is missing?

# Tokenizing

---

Sentences are made up of words. Tokenizing splits up the sentence into its individual words.

At its simplest, tokenizing uses spaces as delimiters between words.

But, sometimes one word is actually a contraction of two:

- ❑ I'm, let's, isn't, won't

CoreNLP tokenizing also splits these into their individual constituents.

Punctuation characters are also tokens

# POS Tagging

---

Once tokenized, the individual tokens can be recognized as **parts of speech**.

Stanford's CoreNLP uses the Penn Treebank Tag Set for recognizing parts of speech.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential <i>there</i>
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP\$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	<i>to</i>
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WPS	Possessive wh-pronoun
36.	WRB	Wh-adverb

# Named Entity Recognition

---

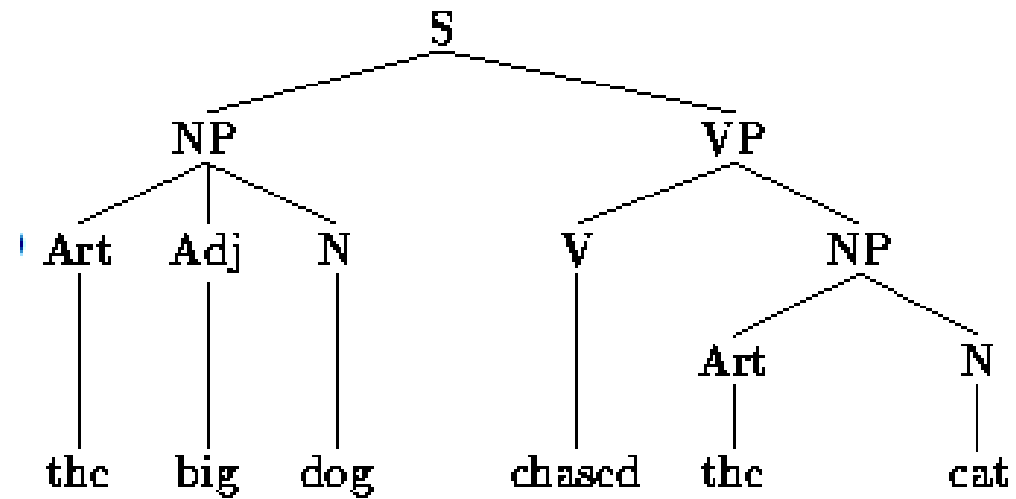
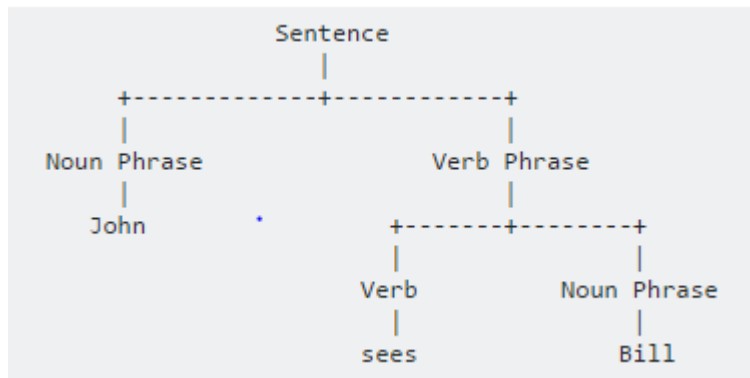
CoreNLP has models and classes for recognizing the names of:

- People
- Places
- Organizations
- Currency
- Time and date
- Extended version: Nationality, Religion, Ideology, Country, State/Province, City, and others.

# Constituency parsing

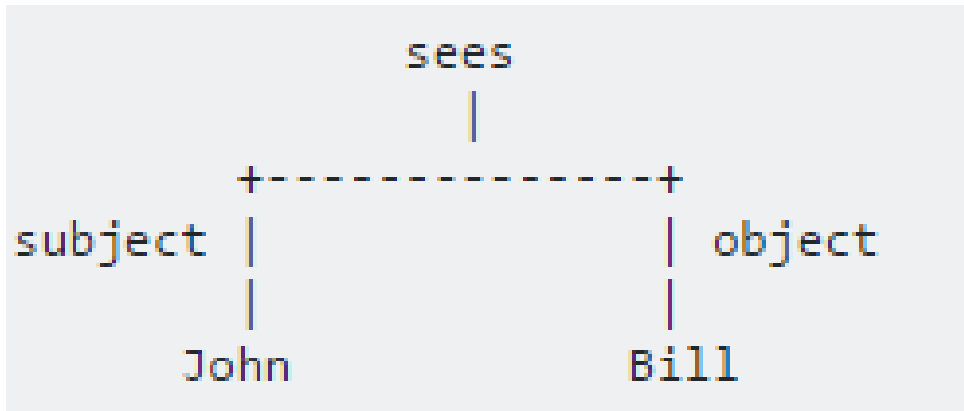
---

Hierarchy of phrases, sub-phrases, etc.



# Dependency parsing

---



Dependency relationships (binary predicates) between words in a sentence.

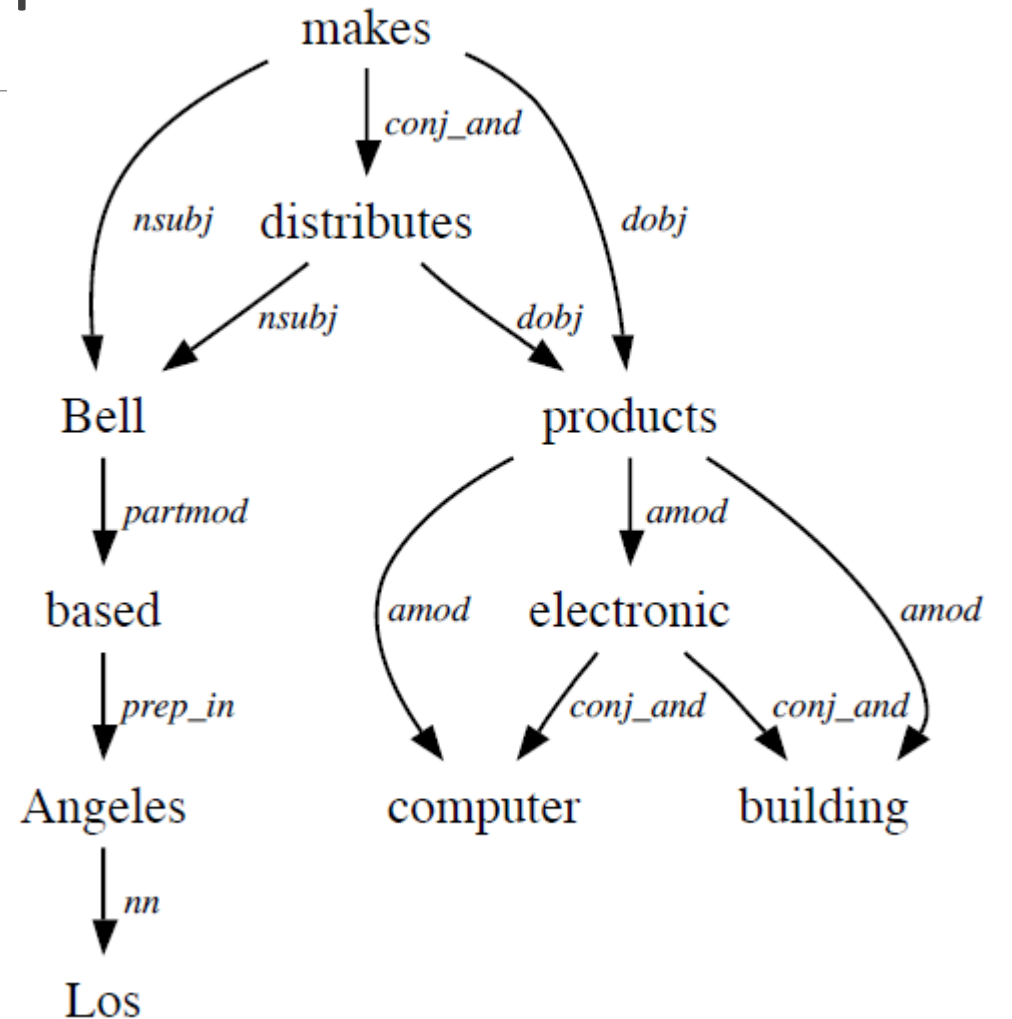
In this picture, the **dependent** points to the **governor**

# Example Dependency Graph

*“Bell, based in Los Angeles, makes and distributes electronic, computer and building products.”*

From Stanford Typed Dependencies Manual (2008)  
[http://nlp.stanford.edu/software/dependencies\\_manual.pdf](http://nlp.stanford.edu/software/dependencies_manual.pdf)

Each connecting line is a dependency relationship. In this figure, the **governor** points to the **dependent**.



# Dependencies from previous graph

---

**nsubj** – nominal subject

**nsubjpass** – nominal passive subject

**dobj** – direct object

**amod** – adjectival modifier

**conj\_and** – conjoint and

**prep** – preposition (e.g. in, on, at, etc.)

**nn** – noun compound modifier

**partmod** – participial verb modifier

# How to recognize subject-object relationships in text?

---

Subject-object relationships -- Who did what to whom?

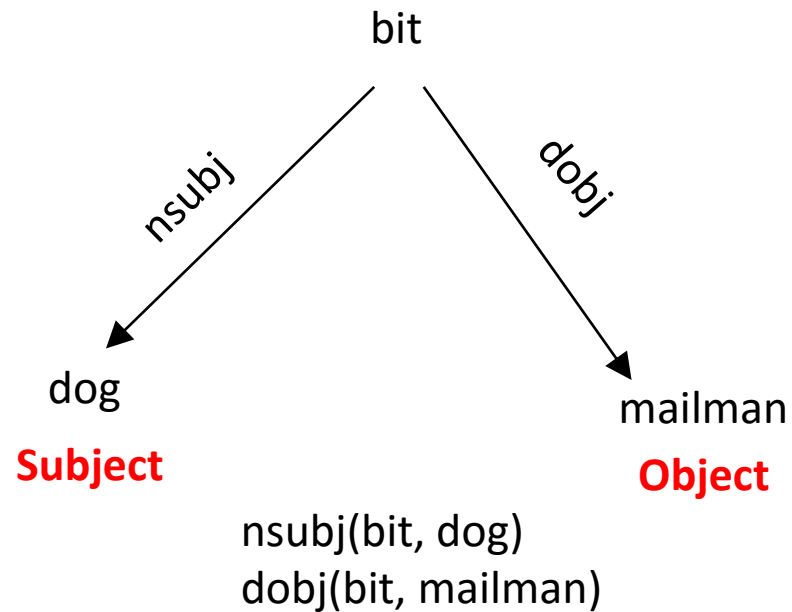
- The dog bit the mailman
- The mailman was bit by the dog
- Trump beat Rubio in Florida, but he was defeated by Kasich in Ohio.
- The sun exerts gravity on the earth and on Mars.
- Hurricane Matthew bashes Florida with 100mph winds

# Subject-object relationships

---

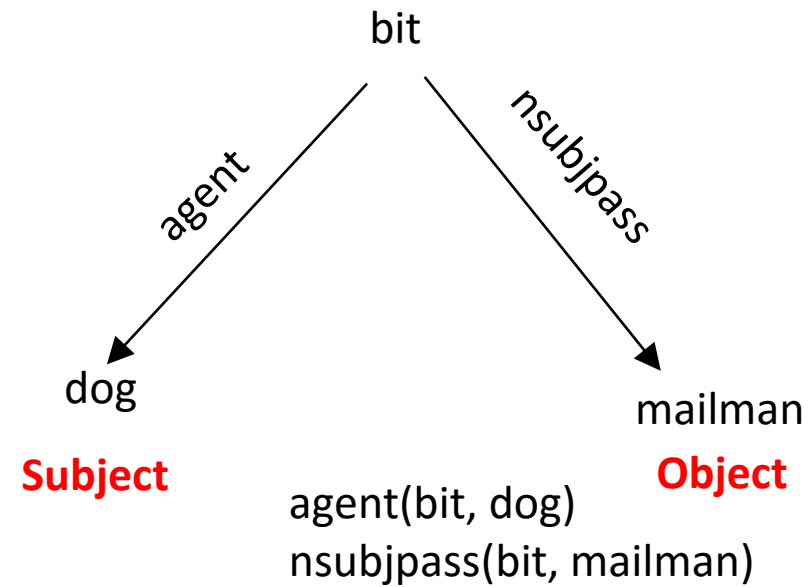
## Active Voice

The dog bit the mailman.



## Passive Voice

The mailman was bit by the dog.



# Key subject-object dependency subgraphs in these sentences

---

Donald J. Trump beat[a] Marco Rubio in Florida, but then he[a] was defeated by John Kasich[a] in Ohio. Later, though, he[b] beat[b] Kasich[b] and all others at the convention.

- nsubj(beat[a],Trump), dobj(beat[a],Rubio)
- nsubjpass(defeated,he[a]), agent(defeated,Kasich[a])
- nsubj(beat[b],he[b]), dobj(beat[b],Kasich[b])

Note: there are six entities (subjects or objects) cited: Trump, Rubio, he[a], Kasich[a], he[b], and Kasich[b]

# Coreference Resolution

---

Identification of **coreference chains**

A coreference chain has a list of **mentions**

Each mention refers to a word (or cluster of words) in the text

- ❑ Mention type – list, nominal, pronominal, proper
- ❑ Gender – male, female, or neutral
- ❑ Animacy – animate or inanimate

# Coreference Resolution

---

Donald J. **Trump** **beat**[a] **Marco Rubio** in Florida, but then **he**[a] **was defeated** by **John Kasich**[a] in Ohio. Later, though, **he**[b] **beat**[b] **Kasich**[b] and all others at the convention.

What are the **coreference chains**?

1. **Trump** and **he**[a] and **he**[b]
2. **Kasich**[a] and **Kasich**[b]

Resulting dependencies after **coreference resolution**. Results in three entities (Trump, Rubio, and Kasich[a]):

- **nsubj(beat[a],Trump), dobj(beat[a],Rubio)**
- **nsubjpass(defeated, Trump), agent(defeated,Kasich[a])**
- **nsubj(beat[b], Trump), dobj(beat[b], Kasich[a])**

# Limitations of NLP

Accuracy measures of CoreNLP annotators.

CoreNLP Annotator	F1 Score	Test Information Source
POS Tagging	97	<a href="https://nlp.stanford.edu/software/pos-tagger-faq.html">https://nlp.stanford.edu/software/pos-tagger-faq.html</a>
Dependency Parsing	81	<a href="https://nlp.stanford.edu/software/stanford-dependencies.shtml">https://nlp.stanford.edu/software/stanford-dependencies.shtml</a>
Named Entity Recognition	81	<a href="https://nlp.stanford.edu/software/crf-faq.shtml">https://nlp.stanford.edu/software/crf-faq.shtml</a>
Coreference Resolution (NN)	60	<a href="https://stanfordnlp.github.io/CoreNLP/coref.html">https://stanfordnlp.github.io/CoreNLP/coref.html</a>

# What is an F1 Score?

---

In machine learning, an F1 score is a metric for measuring accuracy in machine learning (data mining)

Machine learning model will make a prediction. For example prediction of whether a customer will buy a product (simple yes/no). The accuracy of the model is how reliable the prediction is.

F1 is a score between 0 and 100

Two components of F1:

- ❑ **Precision** = What percentage of times did the model predicted yes and the customer actually bought a product.
- ❑ **Recall** = What percentage of times that the customer actually bought a product did the model predict this?

There is often a tradeoff between precision and recall. Think about law enforcement.

# Winograd Schema Challenge (WSC)

---

Hector Levesque – [Winograd Schema Challenge](#)

Common sense reasoning, alternative to the Turing Test

Example: pronoun resolution

“The trophy couldn’t fit in the briefcase because it was too big.”

vs.

“The trophy couldn’t fit in the briefcase because it was too small.”

Coreference resolution’s poor performance shows that computers have a long way to go to pass the WSC.

# Data Visualization for NLP

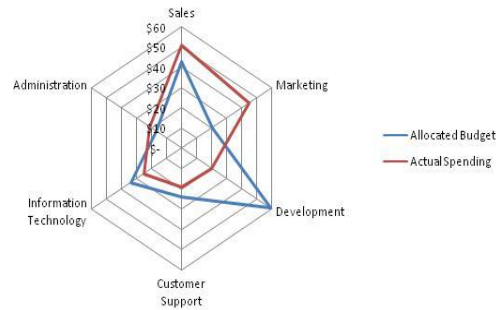
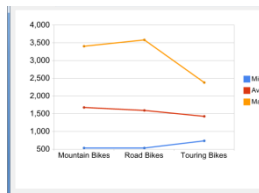
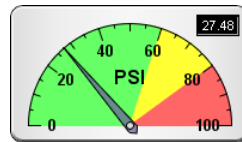
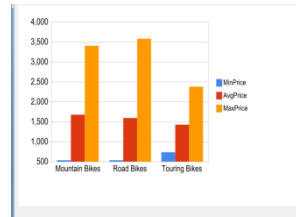
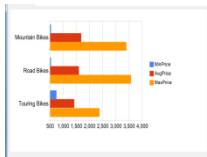
---

To recap, NLP gives us this:

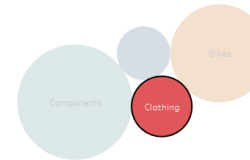
- ❑ Individual sentences with individual tokens (including words)
- ❑ Parts of speech for each word
- ❑ Binary dependencies between pairs of words
- ❑ Name recognition of people, groups, places, times, etc.
- ❑ Coreference chains involving mentions (with a head mention)
- ❑ Story Analyzer creates ASO instances

What visualizations are good for capturing and displaying this data structure and content?

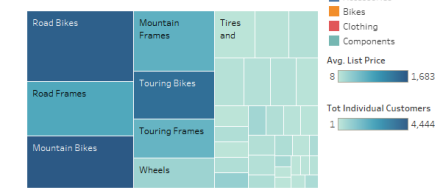
# Common Visualizations found in many Dashboards



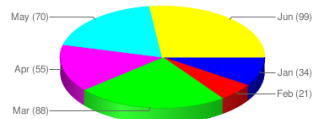
Product Category Bubbles



Subcategory Treemap

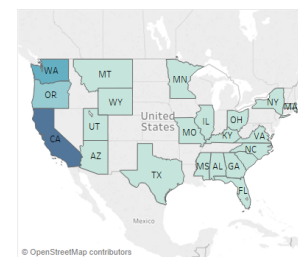


Computer Sales for 1st Half of 2009

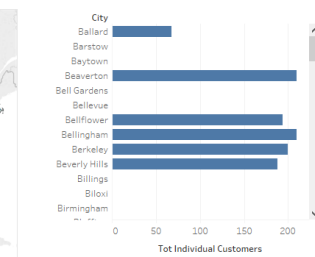


These depict numerical and categorical data – **Structured Data**

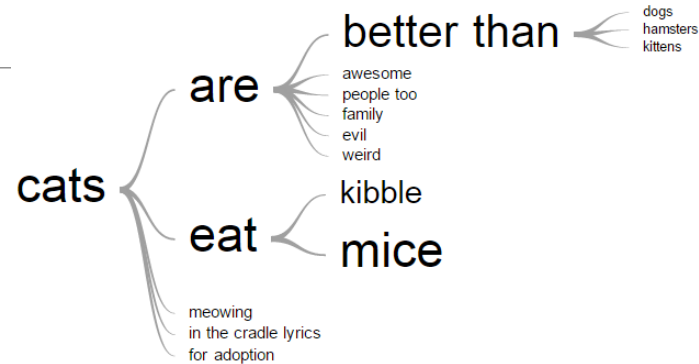
US Map Individual Customers



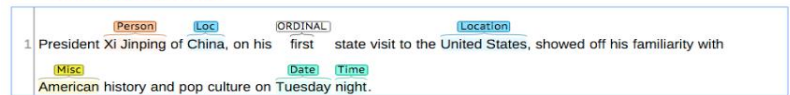
Customers by City



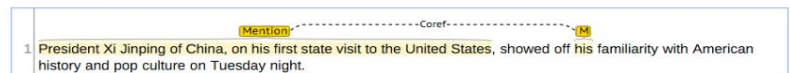
# Visualizations for NLP Results and Concepts



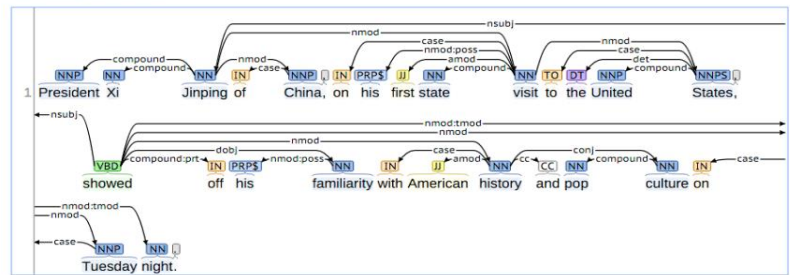
## Named Entity Recognition:



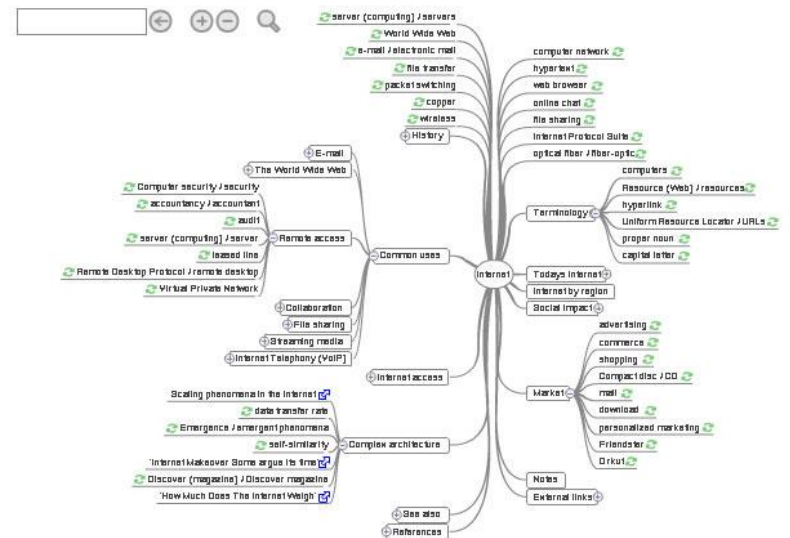
## Coreference:



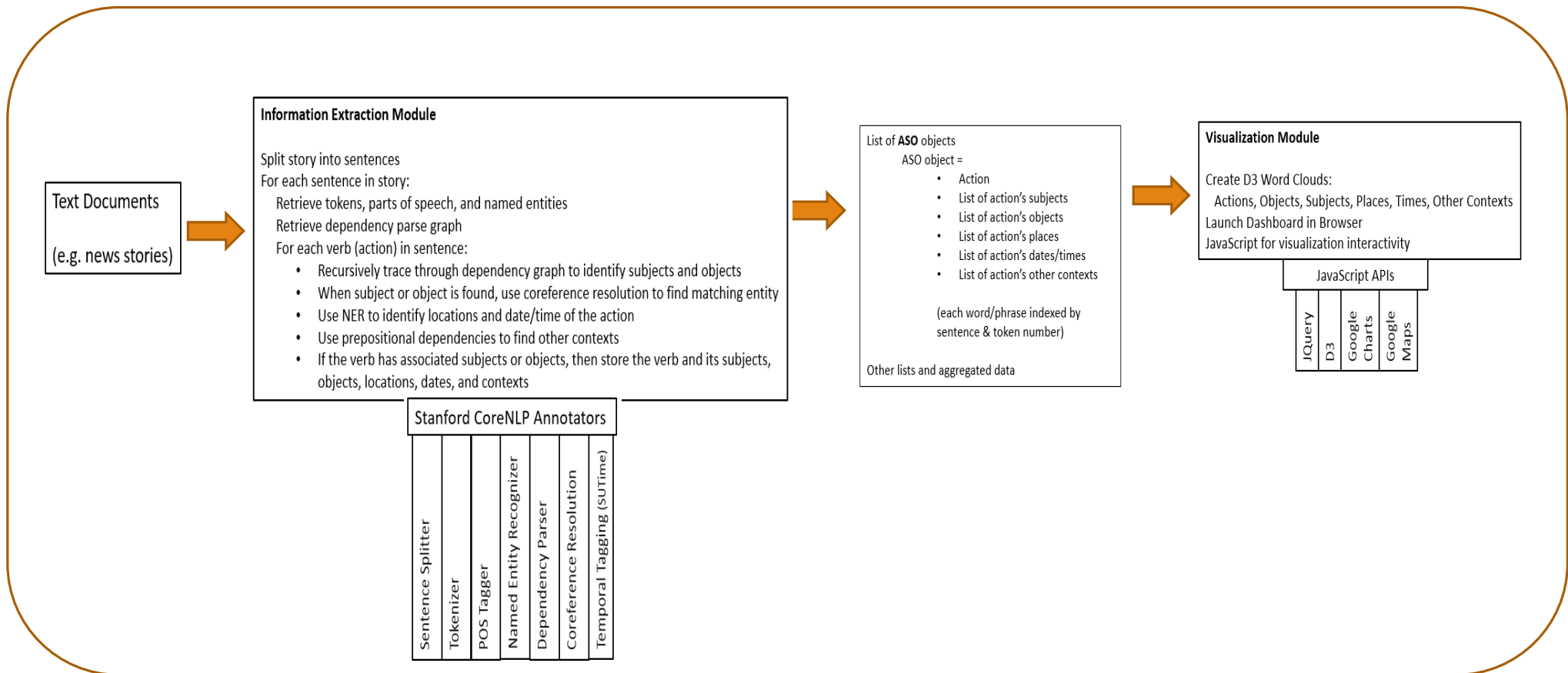
## Basic Dependencies:



NLP takes “unstructured data” (raw text) and imposes a structure



# Story Analyzer Architecture



# Information Extraction

---

Extracting structured information from unstructured text with help from NLP

Often called “slot-filling”

Story Analyzer structure builds on CoreNLP structure, relies heavily on named entity recognition (to recognize people, groups, etc.), dependency parsing (to recognize interactions), and coreference resolution (to identify the head mention)

## Action/Subject/Object (ASO) instances

□ Each action (verb) will have

- List of subjects
- List of objects
- List of places
- List of times
- List of other contexts

# Story Analyzer Dashboards have Six Sections

---

The six sections

Russian Government Links to and Contacts with the Trump Campaign	
A. Campaign Period (September 2015 - November 8, 2016) - 1. Trump Tower Moscow	
Narrative and Highlighted Information	∨
People, Groups, Interactions, and Narrative Web	∨
Dates and Times	∨
Locations	∨
Subjects, Actions, and Objects	∨
Verbs, Nouns, and Contexts	∨

Narrative

IV. RUSSIAN GOVERNMENT LINKS TO AND CONTACTS WITH THE TRUMP CAMPAIGN

The Office identified multiple contacts... links... in the words of the Appointment Order between Trump Campaign officials and individuals with ties to the Russian government.

Highlighted Information

Person Michael Cohen-13-27:

Sentence # 27) From the fall of 2015 until the middle of 2016, Michael Cohen spearheaded the Trump Organization's pursuit of a Trump Tower Moscow project, including by reporting on the project's status to candidate Trump and other executives in the Trump Organization .290

Sentence # 45) In approximately September 2015, Felix Sater, a New York based real estate advisor, contacted Michael Cohen, then-executive vice president of the Trump Organization and special counsel to Donald J. Trump.

Sentence # 49)

Sater contacted Cohen on behalf of I.C. Expert Investment

Visualization APIs:

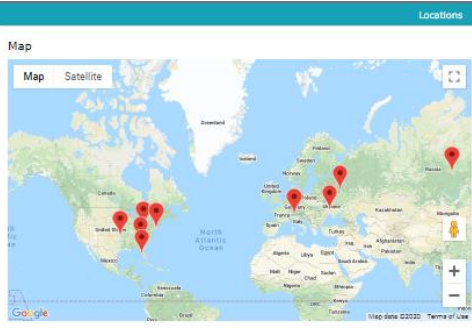
- d3 (data driven documents)
Google

Tables

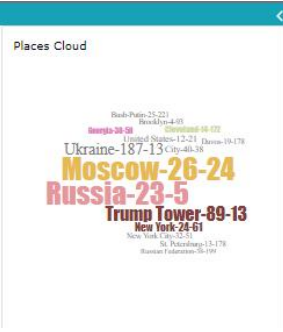
People, Groups, Interactions, Narrative Web. Includes tables for People and Groups, and visualizations for Chord and Force graph.

Force graph

Chord



Map



Interrelated word clouds

Dates and Times visualization including Timeline and Times Cloud.

Timeline

Verbs, Nouns, and Contexts visualization including lists of words and a Contexts Cloud.

Verbs

Nouns

Contexts Cloud

Subjects, Actions, and Objects visualization including Subjects Cloud, Actions Cloud, and Objects Cloud.

Subjects Cloud

Actions Cloud

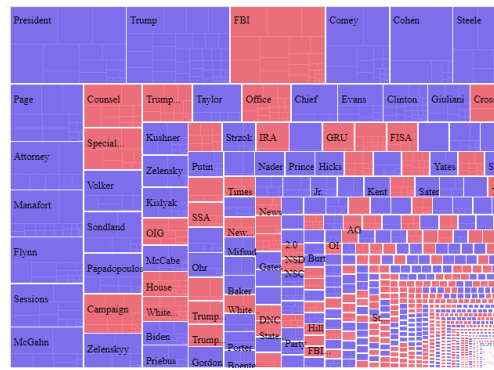
Objects Cloud

# Getting the Big Picture

A separate dashboard:

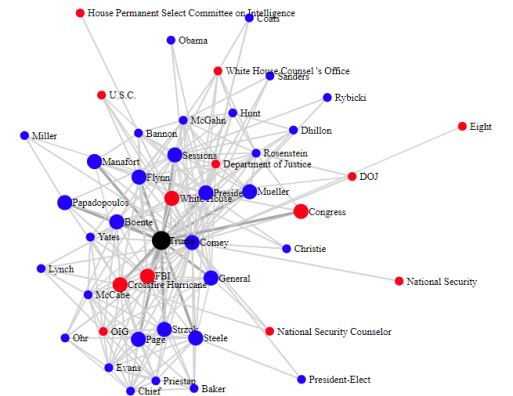
<http://storyanalyzer.org/overviewdashboard.html>

Presents a TreeMap visualization. Can see people and groups, and their interactions, across multiple documents. Clicking a name produces a force graph depicting interactions with other people or groups. Hovering over a node shows documents and mentions. Clicking a mention takes you to the specific dashboard.



## Dashboards with Trump

- alexander vindman testimony mentions: [President Trump-7-54](#)
- darid holmes testimony mentions: [President Trump-31-52](#) [President Trump-29-74](#)
- george kent testimony mentions: [Trump-2-31](#)
- gordon sondland testimony 2 mentions: [President Trump-16-22](#)
- gordon sondland testimony mentions: [Trump-12-102](#)
- horowitz ch1 mentions: [Donald J. Trump-31-4](#)
- horowitz ch3 parts III&IV mentions: [candidate Trump-14-62](#) [Trump-21-300](#)
- horowitz ch5 part I mentions: [Donald J. Trump-41-39](#)
- horowitz ch5 part II A mentions: [Trump-14-24](#)
- horowitz ch5 part II B mentions: [candidate Trump-28-151](#)
- horowitz ch5 part IV mentions: [Trump-26-101](#)
- horowitz ch8 part I mentions: [Donald J. Trump-22-23](#) [candidate Trump-45-84](#)
- horowitz ch8 part II defg mentions: [candidate Trump-65-46](#)
- kant volker testimony mentions: [President Trump-18-49](#)
- marie yovanovitch testimony mentions: [Trump-19-54](#) [Trump-31-102](#)
- mueller report volume 1 part 4 - Manafort mentions: [President Trump-14-16](#) [candidate Trump-11-74](#)
- mueller report volume 1 part 4 - Page mentions: [candidate Trump-11-26](#)
- mueller report volume 1 part 4 - Papadopoulos mentions: [Trump-12-13](#)
- mueller report volume 1 part 4 - Post-election contacts - first part mentions: [President Trump-60-102](#) [President Trump-8-197](#) [Trump-1-2](#)
- mueller report volume 1 part 4 - Post-election contacts - second part mentions: [President Elect Trump-29-72](#)
- mueller report volume 1 part 4 - RNC and post-convention mentions: [candidate Trump-38-1](#)
- mueller report volume 1 part 4 - Simes and CNI mentions: [candidate Trump-11-41](#) [candidate Trump-6-6](#)
- mueller report volume 1 part 4 - Trump Tower Meeting mentions: [Eric Trump-13-71](#) [candidate Trump-22-8](#) [President Trump-8-2](#) [Tranka Trump-27-71](#)



# Applied to several Official Governmental Documents related to Impeachment

---

Mueller Report

House impeachment testimony opening statements

Republican Talking Points

Horowitz report

House Impeachment Report

White House Response

Senate Acquittal Report?

Currently 50+ dashboards. Final estimate for this project: 100+

# Dashboard Generation Procedure

---

Can be done by student assistants:

Clean the text

Run against Story Analyzer for NLP processing

Use Story Analyzer Editor to edit NLP results and correct NLP errors

Run Information Extraction (producing A/S/O and other StoryAnalyzer data)

Run Dashboard data generation (json data)

Copy data to dashboard templates and json repository

# Editing NLP Results

600.4(a), which generally covers efforts to interfere with or obstruct the investigation. President Trump reacted negatively to the Special Counsel's appointment. He told advisors that it was the end of his presidency, sought to have Attorney General Jefferson (Jeff) Sessions unrecuse from the Russia investigation and to have the Special Counsel removed, and engaged in efforts to curtail the Special Counsel's investigation and prevent the disclosure of evidence to it, including through public and private contacts with potential witnesses. Those and related actions are described and analyzed in Volume II of the report.

token	index	sentence	pos	ner	lemma
December	6	98	NNP	DATE	December
declassified	2	106	VBN	O	declassify
delay	12	97	NN	O	delay
deliver	32	54	VB	O	deliver
deliver	6	69	VB	O	deliver
Democratic	14	17	JJ	ORGANIZATION	democratic
Democratic	23	17	JJ	ORGANIZATION	democratic
Democratic	9	75	JJ	IDEOLOGY	democratic
Department	2	21	NNP	ORGANIZATION	Department
Department	2	44	NNP	ORGANIZATION	Department
Department	2	71	NNP	ORGANIZATION	Department
Department	9	81	NNP	ORGANIZATION	Department
Department	2	101	NNP	ORGANIZATION	Department
Department	22	109	NNP	ORGANIZATION	Department
described	16	28	VRB	O	describe
described	40	54	VRB	O	describe
described	6	115	VRB	O	describe

S4)The IRA was based in St. Petersburg, Russia, and received funding from Russian oligarch Yevgeniy Prigozhin and companies he controlled.

Dependencies

- > controlled/VBD (root)
- > based/VBN (dep)
- > IRA/NNP (nsubjpass)
- > The/DT (det)
- > was/VBD (auxpass)
- > Petersburg/NNP (prep\_in)
- > St/NNP (nn)
- > ./ (punct)
- > Russia/NNP (appos)
- > ./ (punct)
- > received/VBD (conj\_and)

XML Dependencies

```
<dependencies style="typed">
<dep type="det">
<governor idx="2">IRA</governor>
<dependent idx="1">The</dependent>
</dep>
<dep type="nsubjpass">
<governor idx="4">based</governor>
<dependent idx="2">IRA</dependent>
</dep>
<dep type="nsubjpass">
<governor idx="12">received</governor>
<dependent idx="6">received</dependent>
</dep>
<dep type="nsubjpass">
<governor idx="12">received</governor>
<dependent idx="12">received</dependent>
</dep>
</dependencies>
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
The	IRA	was	based	in	St.	Peter...	.	Russia	,	and	receiv...	funding	from	Russl...	oligarch	Yevge...	Prigoz...	and	comp...	he	contro...	.
DT	NNP	VBD	VRB	IN	NNP	NNP	.	NNP	,	CC	VRB	NN	IN	JJ	NN	NNP	NNP	CC	NNS	PRP	VRB	.
O	ORGA...	O	O	O	LOCA...	CITY	O	COU...	O	O	O	O	O	NATI...	O	PERS...	PERS...	O	O	O	O	O

Correcting and refining NLP results.

Information extraction and dashboard data generation.

chainID	mentionID	text	sentNum	startIndex	endIndex	mentionType	gender	animacy
Chain: 21	86	5 The Internet Re...	3	1	8	PROPER	NEUTRAL	INANIMATE
Chain: 46	638	10 the United State...	3	34	37	PROPER	NEUTRAL	INANIMATE
Chain: 64	817	13 Russia ,	4	9	10	PROPER	NEUTRAL	INANIMATE
Chain: 86	86	15 The IRA was	4	1	3	PROPER	NEUTRAL	INANIMATE
Chain: 89	21	19 Russian oligarc...	4	15	19	PROPER	MALE	ANIMATE
Chain: 155	21	21 he controlled	4	21	22	PRONOMINAL	MALE	ANIMATE
Chain: 181	675	25 Vladimir Putin	5	11	12	PROPER	MALE	ANIMATE
Chain: 205	86	26 the IRA sent	6	4	6	PROPER	NEUTRAL	INANIMATE
Chain: 206	638	28 the United State...	6	9	12	PROPER	NEUTRAL	INANIMATE
Chain: 213	86	32 The IRA later	7	1	3	PROPER	NEUTRAL	INANIMATE
Chain: 227	46	37 the U.S. political...	7	15	19	NOMINAL	NEUTRAL	INANIMATE
Chain: 265	46	38 it termed	7	21	22	PRONOMINAL	NEUTRAL	INANIMATE
Chain: 265	823	40 Trump and	8	30	31	PROPER	MALE	ANIMATE
Chain: 265	712	41 Clinton .	8	34	35	PROPER	MALE	ANIMATE
Chain: 285	46	46 the U.S. electora	8	15	19	NOMINAL	NEUTRAL	INANIMATE

# Conclusions and Future Research

---

## Applications to other areas:

- Health care – health claim fraud detection
- News – enhanced newsreader experience
- Information Systems discipline – User stories
- Law – Analyzing legal documents
- Business --

## Improvements and enhancements

- Refine information extraction algorithm
- Incorporate ontologies and “common sense”
- Experiment with more visualization techniques